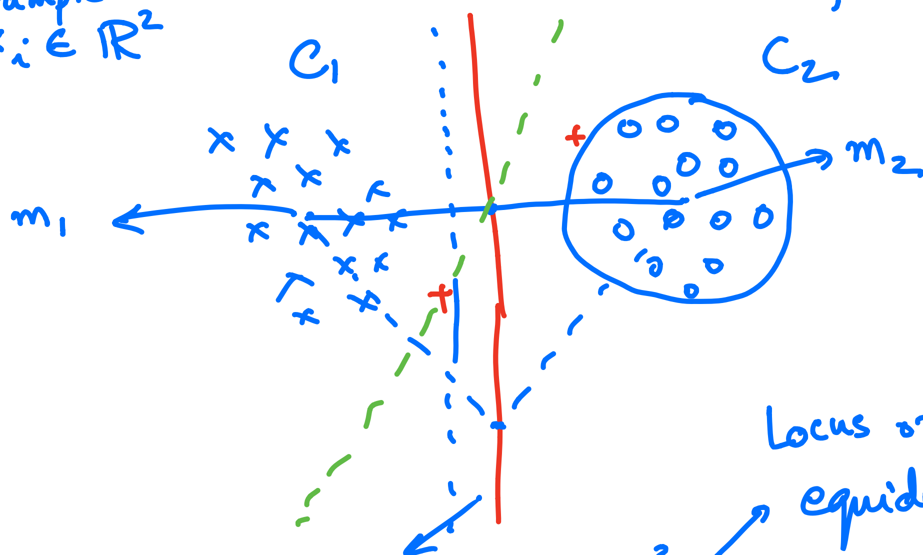
Feb 21, 2020

Classification

Training set : $(x_1, y_1), (x_2, y_2) \ldots\ldots (x_n, y_n)$

Goal : Learn f that predicts class label of new (test) point x.

Example: $x_i \in \mathbb{R}^2$



Locus of points equidistant to m_1 & m_2

$$\|x - m_1\|_2^2 = \|x - m_2\|_2^2$$

$$\cancel{x^T x} - 2x^T m_1 + \|m_1\|_2^2 = \cancel{x^T x} - 2x^T m_2 + \|m_2\|_2^2$$

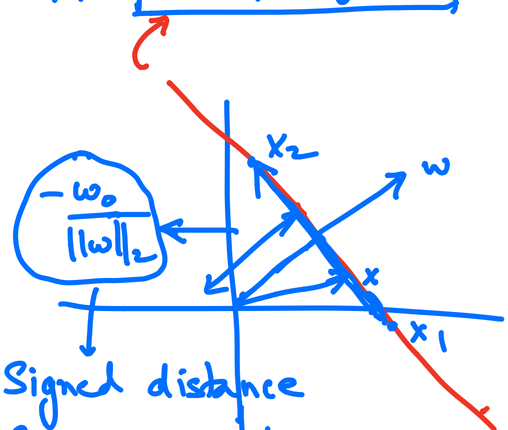
$$\boxed{y(x) = (m_2 - m_1)^T x + \frac{1}{2}\left(\|m_1\|^2 - \|m_2\|^2\right) = 0}$$

$y(x) > 0 \ , \ x \in C_2$

$y(x) < 0 \ , \ x \in C_1$

Example of a hyperplane (linear variety of $(d-1)$ dimensions when d is dimensionality of data)

$$H: \boxed{w^T x + w_0 = 0} \quad \text{- Hyperplane}$$



Signed distance from origin to hyperplane

If x lies on the hyperplane,

$$w^T x + w_0 = 0$$

$$\left(\frac{w}{\|w\|_2}\right)^T x = \frac{-w_0}{\|w\|_2}$$

$$w^T x_1 + w_0 = 0$$

$$w^T x_2 + w_0 = 0$$

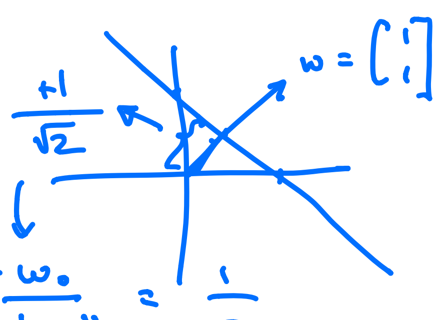
(outward) $w^T(x_1 - x_2) = 0$

w is normal to (the points on) hyperplane

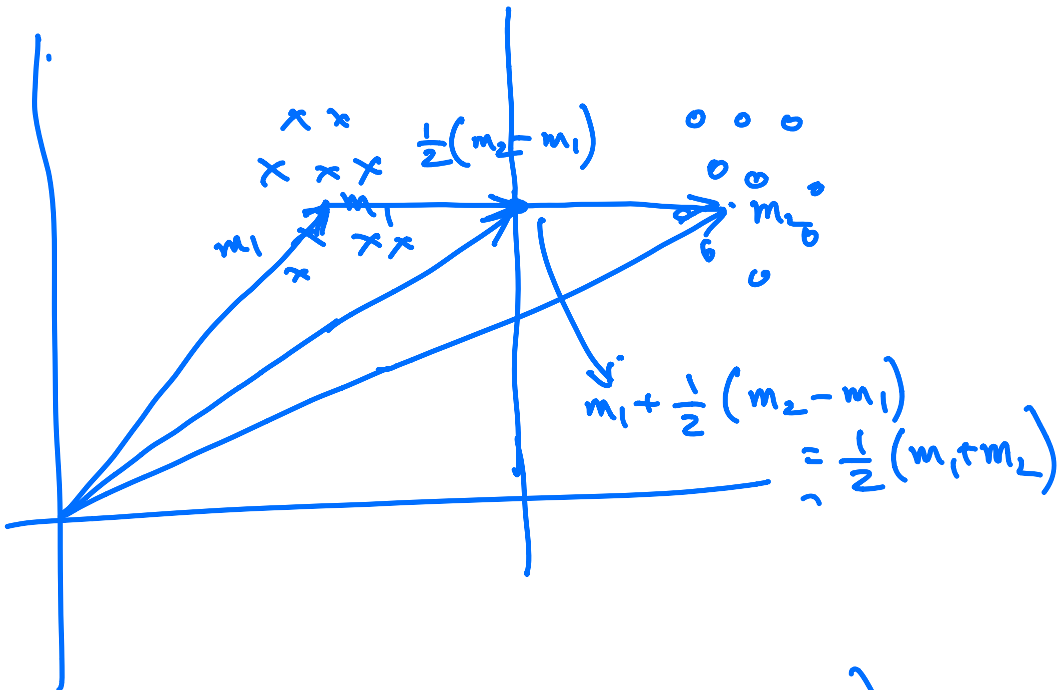
Example: $x_1 + x_2 = 1$

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 1 = 0$$

$$w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad w_0 = -1$$



$$\frac{-w_0}{\|w\|_2} = \frac{1}{\sqrt{2}}$$



$$y(x) = (m_2 - m_1)^T x + \frac{1}{2}\left(m_1^T m_1 - m_2^T m_2\right)$$

$$p(C_1|x) \qquad p(C_2|x)$$

data likelihood

prior

posterior

$$p(C_1|x) = \frac{p(x|C_1)\,p(C_1)}{p(x)}$$ — Bayes Rule

$$p(C_2|x) = \frac{p(x|C_2)\,p(C_2)}{p(x)}$$

MAP rule

Maximum a posteriori probability : $\arg\max_i \, p(C_i|x)$

Gaussian Model: $p(x|C_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$

$$\arg\max_i \log p(C_i|x) = \arg\max_i \log\big(p(x|C_i)\, p(C_i)\big)$$

$$\arg\max_i \boxed{\log p(x|C_i) + \log p(C_i)}$$

$$\boxed{\log p(x|C_i)} = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)$$

Decision Surface: $-\log p(C_1|x) = -\log p(C_2|x)$

$$\boxed{\begin{aligned}&\cancel{\tfrac{d}{2}\log 2\pi} + \tfrac{1}{2}\log|\Sigma_1| + \tfrac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \log p(C_1) = \\ &\quad \cancel{\tfrac{d}{2}\log 2\pi} + \tfrac{1}{2}\log|\Sigma_2| + \tfrac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) - \log p(C_2)\end{aligned}}$$

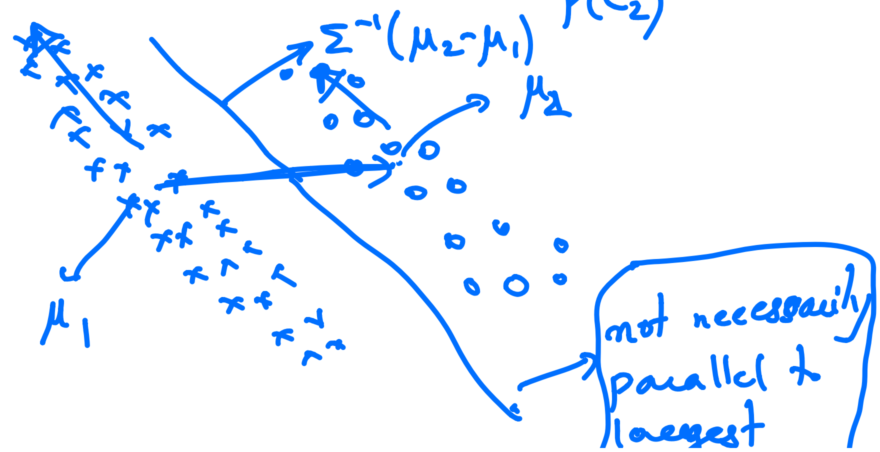
<u>Case I:</u> $\Sigma_1 = \Sigma_2 = I$

$$\frac{1}{2}(x-\mu_1)^T(x-\mu_1) - \log p(C_1) = \frac{1}{2}(x-\mu_2)^T(x-\mu_2) - \log p(C_2)$$

$$\frac{1}{2}\left(\cancel{x^T x} - 2\mu_1^T x + \|\mu_1\|^2\right) - \log p(C_1) =$$

$$\frac{1}{2}\left(\cancel{x^T x} - 2\mu_2^T x + \|\mu_2\|^2\right) - \log p(C_2)$$

$$(\mu_2-\mu_1)^T x + \frac{1}{2}\left(\|\mu_1\|^2 - \|\mu_2\|^2\right) - \log\frac{p(C_1)}{p(C_2)} = 0$$



Case II: $\Sigma_1 = \Sigma_2 = \Sigma$ eigenvector of Σ

$$\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \log p(c_1) = \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) - \log p(c_2)$$

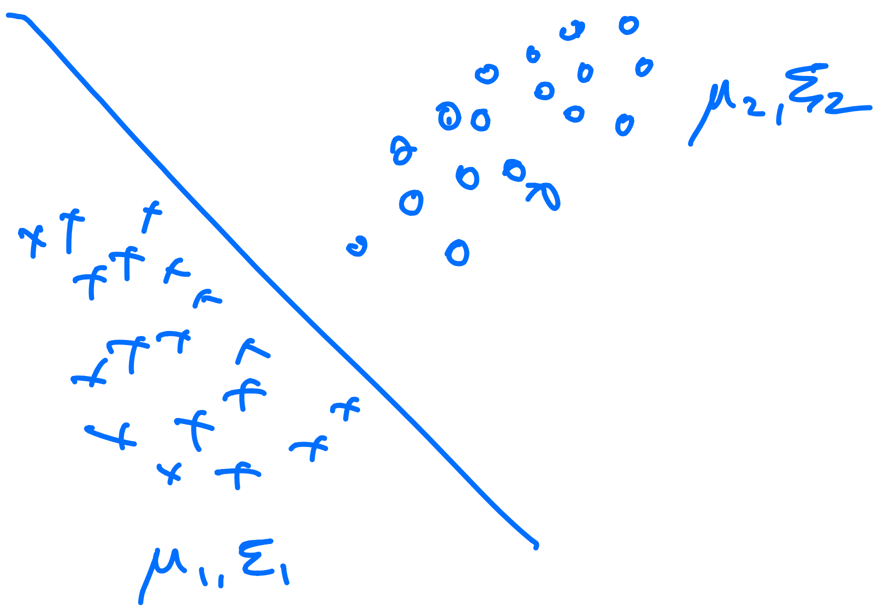
Simplify: $$(\mu_2-\mu_1)^T\Sigma^{-1}x + \frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) - \log\frac{p(c_1)}{p(c_2)} = 0$$

$\downarrow$

Decision Surface is Linear

$$w^T x + w_0 = 0$$

$$w = \Sigma^{-1}(\mu_2-\mu_1)$$



Classification : Regression Approaches

(x_i, y_i), $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}^K$

$i = 1, 2, \dots N$

k-th position

$y_i = [0, 0, \dots, 1, 0 \dots 0]$

if $x_i \in C_k$

$w_0 + w_1 x(1) + w_2 x(2) + \dots + w_d x(d)$ — Linear Fit

Linear Discriminants: $w_k^T x + w_{k0}$ for k-th class

K (K-class problem)

$$\begin{array}{cc} & d+1 \\ \begin{array}{c} C_1 \\ \\ N \\ C_2 \\ \vdots \\ C_K \end{array} & \begin{bmatrix} 1 & x_1(1) & x_1(2) & \dots & x_1(d) \\ & & & & \\ \dots & \dots & \dots & \dots & \dots \\ & & & & \\ \dots & \dots & \dots & \dots & \dots \\ \vdots & & & & \vdots \\ 1 & x_N(1) & x_N(2) & \dots & x_N(d) \end{bmatrix} \end{array} \begin{array}{cc} & K \\ d+1 & \begin{bmatrix} w_{10} & w_{20} & \dots & w_{K0} \\ \underline{w}_1 & \underline{w}_2 & \dots & \underline{w}_K \end{bmatrix} \end{array} \approx \begin{array}{cc} & \\ \begin{array}{c} C_1 \\ \\ \\ \\ C_2 \\ \\ \\ \vdots \\ C_K \end{array} & \begin{bmatrix} 1 & 0 & \cdot & y_1^T & \cdot & 0 \\ 1 & 0 & \cdot & y_2^T & \cdot & 0 \\ 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & \cdot \\ 0 & 1 & 0 & \cdot & \cdot & \cdot \\ & & \vdots & & & \\ & & & y_N^T & & \end{bmatrix} \end{array}$$

$$XW \approx Y \Rightarrow \min_W \frac{1}{2} \|XW - Y\|_F^2$$

$$\equiv \min_W \frac{1}{2} \mathrm{Tr}\left[(XW - Y)^T (XW - Y)\right]$$

$$(X^T X) W^* = X^T Y \Rightarrow W^* = (X^T X)^{-1} X^T Y$$

XW^* is prediction on training data

$x_i^T W$ could be $[\underline{-5 \quad -3 \quad 0 \quad 3 \quad 2 \quad 1 \quad 0.5 \quad 0.1}]$

and this is not a good approximation to 1-hot vectors

Least squares fit has obvious drawbacks

Special case of least squares fit:

2-class problem

$$y_i = \frac{N}{n_1} \text{ for each } x_i \in C_1, \quad n_1 = |C_1|$$

$$y_i = -\frac{N}{n_2} \text{ for each } x_i \in C_2$$

Turns out to be equivalent to:

Fisher's Linear Discriminant

Logistic Regression

We had modeled each class as a Gaussian with covariance Σ:

$$\log \frac{p(C_i|x)}{p(C_j|x)} = \underbrace{\log \frac{p(C_i)}{p(C_j)} - \frac{1}{2}(m_i + m_j)^T \Sigma^{-1}(m_i - m_j)}_{w_0} + \underbrace{x^T \Sigma^{-1}(m_i - m_j)}_{w^T x}$$

K-class problem

$$\log \frac{p(C_1|x)}{p(C_K|x)} = w_{10} + w_1^T x \quad \to \tilde{w}_1^T \tilde{x} \qquad - \text{①}$$

$$\log p(C_2|x) = w_{20} + w_2^T x \qquad - \text{②}$$

$$\vdots \qquad\qquad \vdots$$

$$\log \frac{p(C_{K-1}|x)}{p(C_K|x)} = w_{K-1,0} + w_{K-1}^T x \qquad - \text{(k-1)}$$

Write $p_i = p(C_i|x)$

Add $e^{①}, e^{②}, \ldots e^{(k-1)}$

$$e^{①} \Rightarrow \frac{p_1}{p_K} = e^{w_{10}+w_1^T x}$$

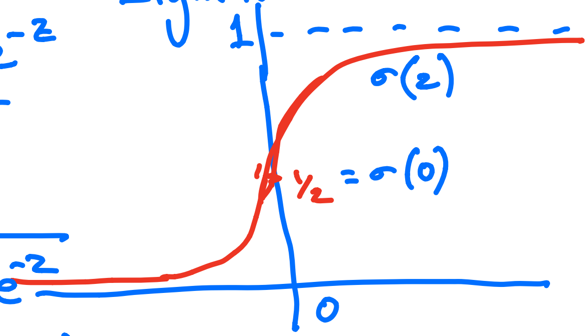
$$\frac{p_1}{p_K} + \frac{p_2}{p_K} + \cdots \frac{p_{K-1}}{p_K} = \sum_{i=1}^{K-1} e^{w_i^T x}$$

$$1 - p_K = p_K \sum_{i=1}^{K-1} e^{w_i^T x}$$

$$\Rightarrow p_K = \frac{1}{1 + \sum_{i=1}^{K-1} e^{w_i^T x}}$$

$$p_i = \frac{e^{w_i^T x}}{1 + \sum_{i=1}^{K-1} e^{w_i^T x}} \quad , i = 1, 2, \ldots k-1$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$ is called the logistic sigmoid function



$$\sigma'(z) = +\frac{1}{(1+e^{-z})^2} \cdot e^{-z}$$

$$= \frac{e^{-z}}{1+e^{-z}} \cdot \frac{1}{1+e^{-z}}$$

$$= (1-\sigma(z)) \cdot \sigma(z)$$

$$\sigma' = \sigma(1-\sigma)$$

Parameters in logistic regression $\{w_{i0}, w_i\}_{i=0}^{k-1}$ usually fit by maximum likelihood using the conditional likelihood given X.

Consider 2-class problem

→ $p(C_1|x) = \boxed{p(w) = p}$

→ $p(C_2|x) = 1-p$

$\boxed{(x_i, y_i) \text{ is training data}}$ $i=1,\dots,N$

Data Likelihood

$$\boxed{\prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}}$$

$p(C_1|x)$ when $x \in C_1$

let $y_i = 1$ when $x_i \in C_1$

$y_i = 0$ when $x_i \in C_2$

Max log-likelihood

$$\boxed{\max_w \ell(w) = \max_w \left[\sum_{i=1}^{N} y_i \log p + (1-y_i)\log(1-p)\right]}$$

$$1-p = \frac{1}{1+e^{w^T x_i}}, \quad p = \frac{e^{w^T x_i}}{1+e^{w^T x_i}} \Rightarrow \log p = w^T x_i - \log\left(1+e^{w^T x_i}\right)$$

$\nabla_w \ell(w) = 0$ - Minimizer will satisfy

$$\nabla_w \log p = x_i - \frac{1}{1+e^{w^T x_i}} \cdot e^{w^T x_i} \cdot x_i$$

$$\nabla_w \log(1-p) = -\frac{e^{w^T x_i}}{1+e^{w^T x_i}} \cdot x_i$$

$$\nabla_w \ell(w) = \sum_{i=1}^{N} \left[y_i x_i \left(1 - \frac{e^{w^T x_i}}{1+e^{w^T x_i}}\right) + (1-y_i) x_i \left(-\frac{e^{w^T x_i}}{1+e^{w^T x_i}}\right)\right]$$

$$= \sum x_i \left[y_i - \cancel{\frac{y_i e^{w^T x_i}}{1+e^{w^T x_i}}} - \frac{e^{w^T x_i}}{1+e^{w^T x_i}} + \cancel{\frac{y_i e^{w^T x_i}}{1+e^{w^T x_i}}}\right]$$

$$= \boxed{\sum_{i=1}^{N} x_i \left[y_i - \underbrace{\frac{e^{w^T x_i}}{1+e^{w^T x_i}}}_{p(C_1|x_i)} \right] = 0}$$

logistic regression parameters satisfy this

$$= \sum_{i=1}^{N} x_i \left[y_i - \left(1 - \sigma(w^T x_i)\right) \right]$$

$d+1$ non-linear equations in w
$\downarrow$
$(d+1)$ parameters

Iterated Re-weighted Least squares method (IRLS)

~~*~~ Gradient Descent/Ascent

Newton's Method

Drawback is ~~that~~ each step of Gradient Descent is $O(N)$

Stochastic Gradient Descent

Regularization

$$\lambda \|w\|_2^2$$

$$\lambda \|w\|_1$$